Perbandingan Akurasi Metode *Naïve Bayes Classifier* dan *Random Forest* Menggunakan Reduksi Dimensi *Linear Discriminant Analysis* (LDA) untuk Diagnosisi Penyakit Diabetes

Arif Ariwikri[1], Gatot Aria P.[2], Hardian Fathurahman[3], M. Alfurqon S.F[4], M. Fahreza A.[5]
[1] Fakultas Teknik, Teknik Elektro
Universitas Sriwijaya
Palembang, Indonesia
Penulis korespodensi: *fathurahman0502@gmail.com*

***Abstrak* – Diabetes merupakan suatu penyakit yang dapat menyerang orang – orang di belahan dunia. Jumlah kematian akibat diabetes meningkat dari tahun ke tahun. Diabetes terjadi ketika tubuh tidak menghasilkan cukup insulin. Penyakit ini merupakan penyakit kompleks dan fatal yang memerlukan perawatan medis berkelanjutan untuk menghindari risiko komplikasi. Menganalisis pasien diabetes sejak usia dini bisa memberikan catatan penyakit yang luas dan memungkinkan pencegahan. Salah satu cara untuk melakukannya dengan klasifikasi data mining. Teknik ini dipakai sebagai prediksi siapa yang terkena diabetes dan siapa yang tidak terserang diabetes. Dalam penelitian menggunakan metode Naive Bayes Classifier (NBC) dan Random Forest menggunakan reduksi dimensi LDA. Hasil penelitian menunjukkan akurasi 77,67% untuk algoritma Random Forest dan 76% untuk Naive Bayes Classifier (NBC). Random Forest lebih baik/akurat dibandingkan menggunakan Naive Bayes Classifier dalam mengklasifikasikan diabetes.**

Kata Kunci–Diabetes, Klasifikasi Naive Bayes, Analisis Diskriminan Linier, Random Forest.

***Abstract*— Diabetes is a disease that can affect people all over the world. The number of deaths due to diabetes is increasing year by year. Diabetes occurs when the body doesn't produce enough insulin. The disease is a complex and fatal disease that requires ongoing medical care to avoid the risk of complications. Analysis of diabetic patients from an early age provides a history of a wide range of diseases and enables prevention. One way to do this is classification by data mining. This technology is used to predict who has diabetes and who does not. A random forest method with naive Bayes classifier (NBC) and LDA dimensionality reduction was used in this study. The results showed an accuracy of 77,67% for the random forest algorithm and 76% for the Naive Bayes classifier (NBC). Random Forest is better/more accurate than using Naive Bayes Classifier for diabetes classification.**

Keywords—Diabetes, Naive Bayes Classification, Linear Discriminant Analysis, Random Forest.

I. PENDAHULUAN

Salah satu penyakit keturunan mematikan yang dapat menyerang manusia di belahan dunia adalah diabetes. Berdasarkan data dari WHO (Organisasi Kesehatan Dunia), jumlah orang yang terserang diabetes saat ini mendekati 350 juta. Di tahun 2012, kurang lebih 1,5 juta orang dilaporkan meninggal karena diabetes dan lebih dari 80% kematian di negara berkembang [1]. Penderita diabetes setiap tahun semakin meningkat. Lalu di tahun 2015 terdapat 10 juta orang di Indonesia terserang diabetes. Data tersebut diperoleh dari IDF (*International Diabetes Federation*), para penderita diabetes di Indonesia diproyeksikan meningkat ke 16,2 juta pada tahun 2040. Deteksi dini penyakit diabetes sangat diperlukan untuk mengatasi masalah ini. Deteksi dini diharapkan bisa mengurangi risiko komplikasi untuk para penderita diabetes di kemudian hari.

Deteksi dini seharusnya mengurangi risiko komplikasi di kemudian hari pada penderita diabetes. Menganalisis pasien diabetes secepatnya akan memungkinkan penyakit didokumentasikan secara luas dan dapat dicegah. Hal tersebut bisa dilakukan dengan menggunakan *classifier* pada mining data. Tahap memilih data yang cukup besar lalu disimpan dalam *repository* dan menerapkan pengenalan pola, matematika dan teknik statiska untuk mendapatkan korelasi dan pola yang bermakna disebut Data Mining. Data mining berdampak pada banyak bidang, termasuk dunia kesehatan, dan membantu mendeteksi penyakit [2]. Data mining Ini merujuk pada pencarian kumpulan informasi dibutuhkan dari berbagai data besar di beberapa bidang seperti kedokteran, pendidikan, dan transaksi bisnis. Algoritma tersebut bisa menganalisa data dalam jumlah besar di berbagai bidang, yang paling penting pada bidang kedokteran atau medis. Ini memberikan alternatif untuk pendekatan pemodelan prediksi komputasi rutin, mengurangi kesalahan antara hasil yang diprediksi dan aktual, dan memahami interaksi nonlinier yang kompleks antara berbagai faktor. Oleh karena itu, penelitian ini menggunakan ilmu data mining, khususnya klasifikasi diabetes.

Berdasarkan uraian masalah di atas, Teknik yang dapat dipakai adalah data mining. Analisa dan diagnosa pasien diabetes membutuhkan sebuah teknik yang paling akurat. Oleh karena itu, penelitian ini secara efektif membandingkan beberapa teknik data mining, yaitu klasifikasi Naive Bayes dengan Analisis Diskriminan Linier. Alasan penelitian ini menggunakan dua metode tersebut karena dua metode ini merupakan metode yang paling umum digunakan dalam melakukan penlitian dengan tujuan penelitian untuk membandingkan kedua metode tersebut serta mencari metode mana yang memiliki akurasi yang lebih baik.

II. STUDI PUSTAKA

A. Data Mining

Proses pengolahan atau mengekstrak pengetahuan yang diinginkan dari berbagai data besar atau dikenal juga dengan data mining. Mining data, basis data dan penemuan pengetahuan (KDD) menggambarkan proses pencarian informasi yang tersembunyi dari kumpulan data besar. Meskipun istilah tersebut secara konseptual tidak sama, namun sebenarnya terkait fase dari proses KDD merupakan mining data. Knowledge Discovery in Database (KDD) merupakan eksplorasi sejumlah besar informasi atau pengetahuan yang dibutuhkab. [3]. Teknik penggalian informasi yang tidak didokumentasikan dalam dataset disebut data mining. Sejak tahun 1990, istilah data mining menjadi populer. Data mining adalah suatu keharusan, dan pemrosesan data harus nyaman dan sangat berbeda dari sains, kesehatan, akademisi, atau bahkan industri [4].

B. Naïve Bayes Classifier

Klasifikasi Naive Bayes merupakan suatu proses klasifikasi probabilistik dengan teorema Bayes, menganggap bahwa tiap *feature* mempunyai kinerja sama pada kelas target. Klasifikasi Naive Bayes, yang termasuk dalam kelas tertentu yang berkontribusi pada pengambilan sampel probabilistik, mudah diterapkan, cepat dihitung, dan cocok untuk kumpulan data berdimensi tinggi yang besar. Cocok untuk aplikasi real-time dan tahan terhadap gangguan [5]. Kalsifikasi naïve Bayes mengolah kumpulan data *trainer* untuk menjumlahkan probabilitas kelas dan probabilitas bersyarat yang dapat menentukan berbagai frekuensi setiap nilai *feature* untuk nilai *class* tertentu. Pengklasifikasi Naive Bayes bekerja paling baik ketika mereka berkorelasi. Dikarenakan fitur korelasi dipakai dua kali dalam model, fitur tersebut disembunyikan, yang terlalu menekankan dibutuhkannya fitur yang berkorelasi [6].

Dalam persamaan (1) dan (2), teorema Bayes memudahkan untuk menghitung probabilitas posterior dari P(c|x) dari P(c), P(x), dan P(x|c). klasifikasi dari naive bayes menganggap efek nilai prediktor (x) pada *class* tertentu (c) tidak bergantung prediktor yang lain. Anggapan ini dikenal dengan independensi *class* bersyarat.

$$P(c|x) = \frac{P(\boldsymbol{x}|\boldsymbol{C})\,P(c)}{P(x)} \tag{1}$$

$$P(c|x) = P(x_1|c) \times P(x_1|c) \times P(x_2|c) \times \ldots \times (x_n|c) \times P(c) \tag{2}$$

dimana P(c) merupakan probabilitas kelas sebelumnya, P(x) merupakan probabilitas sebelumnya dari *predictor*, P(c|x) merupakan probabilitas posterior kelas (target)diberikan prediktor (atribut), dan P(x|c) merupakan peluang yang merupakan peluang kelas yang diberikan oleh prediktor.

C. Linear Discriminant Analysis (LDA)

Teknik yang digunakan untuk *pattern recognition* dalam komputasi statistika dengan menemukan hubungan data linier yang memaksimalkan jarak antar kelas dan meminimalkan jarak antara data serupa disebut dengan LDA. LDA menjadi *basic* untuk mengklasifikasikan data. Anda dapat mereferensikan kelas yang ada dari data yang ada dan menampilkan data sampel sebagai {x1, x2, ..., XNi}. Setiap kolom dalam matriks ini mewakili sampel data. Selanjutnya, dengan menggunakan contoh proyeksi matriks X, carilah hasil transformasi Y dari X ke dimensi hyperplane K – 1.[7]. Perhitungan LDA ditunjukkan pada Persamaan 3:

$$fi = \mu_i C^{-1} x_k^T - \frac{1}{2}\mu_i C^{-1} \mu_i^T ln(P_i) \tag{3}$$

Keterangan:

fi = Fungsi diskriminan kelas ke – i

μ_i^T = Transpos rata – rata kelas - i

μ_i = Rata-rata nilai setiap kelas

x_k^T = Transpos dari matriks data uji

C^{-1} = Invers dari grup matriks kovarian

P_i = Peluang munculnya kelas ke – i

D. Random Forest

Jenis prosedur pemecahan penjabaran yg terdiri berdasarkan beberapa pohon keputusan atau Random Forest. Algoritma ini dibentuk sesuai dengan nilai vektor acak yang diambil sampelnya dengan independen dan seragam di semua pohon. [8]. Random Forest adalah salah satu metode *classifier* paling akurat yang digunakan saat membuat prediksi, dapat menangani sejumlah besar variabel input tanpa overfitting, dan dapat mendeteksi perbandingan antar random forest seperti properti metode ansambel [9].

Keuntungan menggunakan algoritma *Random Forest* sebagai metode klasifikasi adalah tidak menimbulkan masalah overfitting dalam *classifier*, Random Forest biasa dipakai untuk klasifikasi Identifikasi fitur utama yang akan digunakan dari kumpulan data pelatihan dan regresi [10].

III. METODE PENELITIAN

Dalam penelitian terdapat reduksi dimensi dengan metode LDA dalam klasifikasi diabetes untuk membandingkan naive bayes *classifier* dengan random forest. Penelitian ini, akan menghasilkan akurasi terbaik antara menggunakan Naive Bayes dengan Random Forest.

A. Sumber Data

Sumber dataset pada penelitian kali ini diperoleh dari webstite Kaggle di alamat web https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset [11]. Dataset yang digunakan pada penelitian ini adalah *Diabetes Patients Data.* Kumpulan data ini didapat dari Institut Nasional Diabetes dan Penyakit Pencernaan dan Ginjal

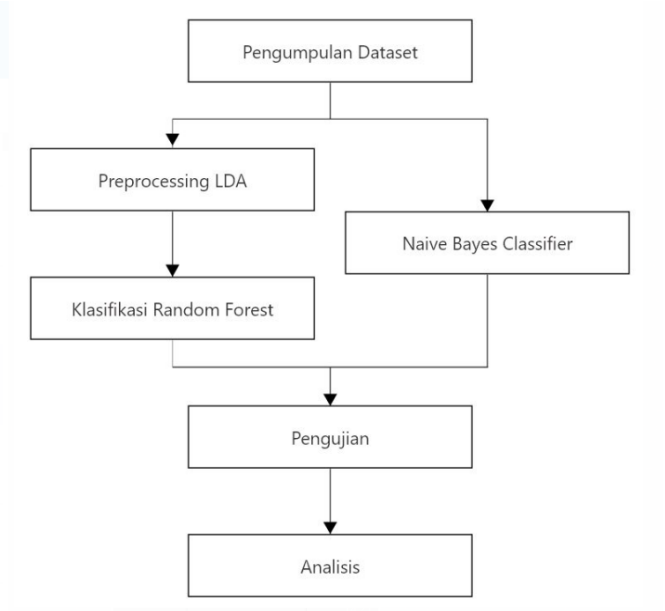
Dalam penelitian ini, 9 variabel dengan total 768 data dan kelas yang ditetapkan digunakan sebagai variabel. Fungsi dari dataset ini adalah sebagai prediksi secara diagnosa, apakah seorang pasien tersebut terserang diabetes yang berdasarkan diagnosa tertentu yang terdapat dalam dataset. Ada batasan dikenakan pada pemilihan sampel ini dari database besar. Khususnya, semua pasien harus berumur minimal 21 tahun dan perempuan Pima India.

B. Split data otomatis

Tahap split data otomatis, 768 data diabetes dipisah menjadi 3 kelompok, data train dengan persentase 80% dan data testing 20%. Lalu, 67% untuk training dan 33% untuk testing, lalu 50% untuk training dan 50% data testing. Data training berfungsi membentuk pola, sedangkan data testing berfungsi untuk model penguji.

C. Metode yang Diusulkan

Klasifikasi yang dipakai adalah Naive Bayes Classifier dan Random Forest menggunakan reduksi LDA. Beberapa tahapan terdapat pada Gambar 1.



Gambar 1. Tahapan-Tahapan Penelitian

D. Pengumpulan Dataset

Tahapan Pertama kali yang dilakukan adalah mencari dan mengumpulkan berbagai dataset dari Pima Indian dari situs web Kaggle. Dataset Pima Indian Diabetes berisi 9 variabel yang dibagi menjadi 8 fitur dan 1 kelas, dengan total 768 data. Detail atribut terdapat pada Tabel 1 dan 2.

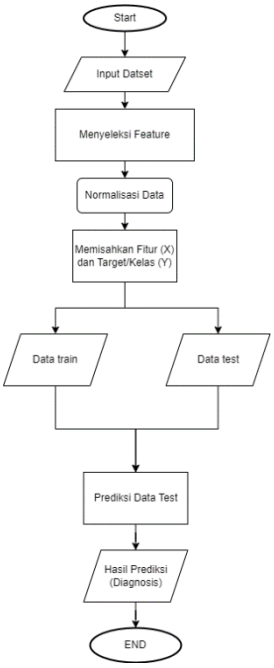
TABLE I. FITUR DATASET

No	Feature	
	Attribute	***Description***
1	Pregnancies	Menunjukkan Nilai dari Kehamilan
2	Glucose	Menunjukkan Level Glukosa pada Darah
3	BloodPressure	Menunjukkan Nilai dari pengukuran tekanan darah
4	SkinThickness	Menunjukkan ketebalan dari kulit
5	Insulin	Menunjukkan level dari insulin pada darah
6	BMI	Menunjukkan Berat Badan
7	DiabetesPedigreeFunction	Untuk Menunjukkan Persentasi dari Diabetes
8	Age	Menunjukkan Umur

TABLE II. KELAS DATASET

No	Class	
	Attribute	***Description***
1	Outcome	Untuk menampilkan nilai akhir yaitu 1 untuk Yes dan 0 untuk No

E. Rancangan Algoritma Sistem



Gambar 2. Flowchart Rancangan Algoritma Sistem

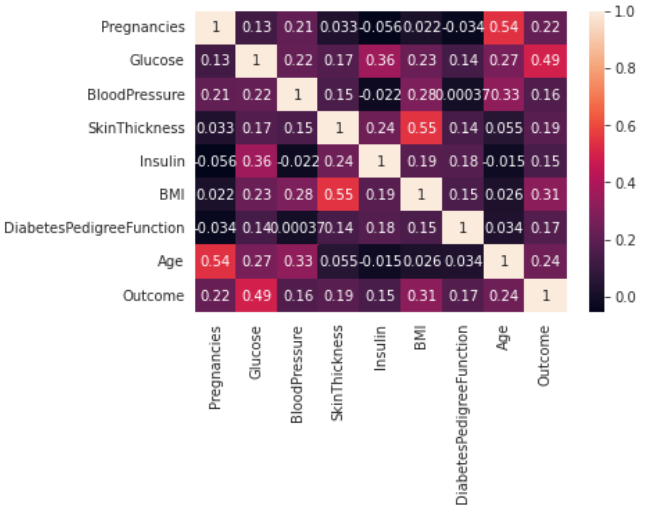
Perancangan skema sistem prediksi bertujuan untuk membuat berupa sistem prediksi penyakit jantung diagnosa pasien. berdasarkan Gambar 2 pada *flowchart* rancangan algoritma sistem. Dimulai input *dataset* yang dipisahkan dari fitur (X) dan target (Y). Lalu normalisasi data, data test dan data train metode dilakukan dengan menggunakan Gaussian Naïve Bayes dan Random Forest, selanjutnya adalah prediksi data test untuk mendapatkan diagnosis atau prediksi hasil.

IV. HASIL DAN PEMBAHASAN

A. *Feature Selection*

Pada melakukan feature selection dimana kita akan mengurangin feature yang akan kita gunakan sebagai perbandingan untuk memprediksi hasil yang akan kita hasilkan. Pada feature selection ini kita memanfaatkan *Pearson Correllation Coefficient*. *Pearson Correlation Coefficient* ini kita gunakan untuk Membantu Anda mengetahui hubungan antara dua besaran. Ini memberi Anda ukuran kekuatan hubungan antara dua variabel. Nilai Koefisien Korelasi Pearson dapat berkisar antara -1 sampai +1. 1 berarti sangat berkorelasi dan 0 berarti tidak ada korelasi.

Pada tahap ini kita akan menampilkan *value* dari Pearson Correlation Coefficient menggunakan heatmap dari Library dari seaborn untuk bentuk data yang terdapat pada Gambar 3.



Gambar 3. Heatmap dari Feature

Berdasarkan Gambar 3 dapat kita lihat bahwa Glucose, BMI dan Umur sangat berkorelasi dengan hasil pada outcome. Pada Gambar 3 juga dapat kita lihat bahwa *BloodPressure, Insulin*, dan *DiabetesPedigreeFunction* tidak memberikan pengaruh yang besar pada data kita sehingga kita dapat menghapusnya.

B. *Preprocessing*

Tahap ini merupakan tahap kritis dalam proses machine learning. Fungsi dari tahapan ini adalah untuk menyatakan bahwa semua variabel memiliki range nilai sama persis dan tidak ada nilai data yang terlalu besar. Fase ini menggunakan *Quantile Transformer* untuk melakukan preprocess data dengan *Naive Bayes Classification*, sedangkan *Random Forest Classification* menggunakan LDA untuk mereduksi data yang digunakan. Hasil *preprocessing* dapat dilihat pada Tabel 3.

TABLE III. HASIL *PREPROCESSING*

No	Pregancies	Glucose	Skin Thickness	BMI	Age	Outcome
1	0.747718	0.810300	0.801852	0.591265	0.88983	1.0
2	0.232725	0.91265	0.644720	0.213168	0.558670	0.0
3	0.863755	0.956975	0.357888	0.077575	0.585398	1.0
4	0.232725	0.124511	0.357888	0.284224	0.000000	0.0
5	0.000000	0.721643	0.801825	0.926988	0.606258	1.0

C. *Memisahkan Fitur (X) dan Target/Kelas (Y)*

Pada Gambar 2 dapat dilihat data yang digunakan dipisahkan menjadi Fitur (X) dan Target/Kelas (Y). Berdasarkan Tabel 3, Fitur (X) dan Target/Kelas (Y) dapat ditentukan seperti yang terdapat di Tabel 4 dan 5.

TABLE IV. FITUR (X)

No	Pregancies	Glucose	Skin Thickness	BMI	Age
1	0.747718	0.810300	0.801852	0.591265	0.88983
2	0.232725	0.91265	0.644720	0.213168	0.558670
3	0.863755	0.956975	0.357888	0.077575	0.585398
4	0.232725	0.124511	0.357888	0.284224	0.000000
5	0.000000	0.721643	0.801825	0.926988	0.606258

TABLE V. TARGET/KELAS (Y)

No	Outcome
1	1.0
2	0.0
3	1.0
4	0.0
5	1.0

D. Pemisahan Data

Data yang dipakai dibagi menjadi dua bagian yaitu data uji dan data latih yang memperlihatkan korelasi jumlah dari data latih, data tes dan presentase yang terlihat pada Tabel 6.

TABLE VI. PEMISAHAN DATA

No	Split data Training dan Testing	Jumlah Data	
		Data Train	Data Test
1	***Training 80% dan Testing 20%***	614	154
2	***Training 67% dan Testing 33%***	514	254
3	***Training 50% dan Testing 50%***	384	384

E. Hasil Prediksi

Hasil prediksi ini adalah hasil yang didapat dari split data menggunakan *Naïve Bayes Classifier* dan *Random Forest*. Pada Tabel 7 dapat dilihat hasill *split data* menggunakan *Naïve Bayes Classifier* dan pada Tabel 8 dapat dilihat *split data* menggunakan *Random Forest*. Dari hasil prediksi kita dapat memperoleh nilai akurasi, F1, dan presisi seperti yang terlihat di Tabel 7 dan 8. Nilai akurasi diperoleh dari persentase jumlah prediksi benar (Diabetes dan Tidak diabetes) dibanding dengan jumlah data test secara keseluruhan. Untuk F1 didapatkan dari nilai rata rata antara nilai dari presisi dan nilai recall dimana nilai recall merupakan perbandingan antara jumlah diprediksikan bernilai positif dengan banyak data positif yang memang benar positif. Terakhir, untuk nilai presisi diperoleh dari jumlah prediksi benar positif dibandingkan dengan jumlah hasil yang diprediksi positif.

1) Naive Bayes Classifier

TABLE VII. SPLIT DATA NAÏVE BAYES CLASSIFIER

No	Split data Training dan Testing	Naive Bayes Classifier		
		Akurasi	F1	Presisi
1	***Training 80% dan Testing 20%***	77%	58,139%	64,102%
2	***Training 67% dan Testing 33%***	75%	57,894%	64,705%
3	***Training 50% dan Testing 50%***	76%	60,085%	68,627%

2) Random Forest

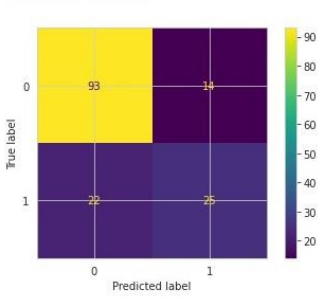
TABLE VIII. RANDOM FOREST

No	Split data Training dan Testing	Random Forest		
		Akurasi	F1	Presisi
1	***Training 80% dan Testing 20%***	79%	59,999%	72,727%
2	***Training 67% dan Testing 33%***	76%	56,338%	68,965%
3	***Training 50% dan Testing 50%***	78%	64,705%	71,962%

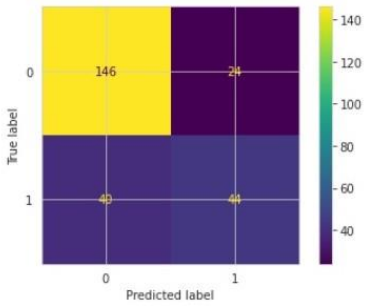
F. Confusion Matrix

Confusion Matrix merupakan tabel yang membandingkan hasil prediksi dan aktual. Matrix *confusion* dapat digunakan untuk mengevaluasi kinerja model klasifikasi. Gambar 2 hingga 6 menunjukkan *matrix confusion* untuk *naïve bayes classifier* untuk perbagian data *training* yang berbeda-beda. Sedangkan Gambar 7 hingga 9 merupakan *confusion matrix* untuk perbagian data *training* yang berbeda-beda

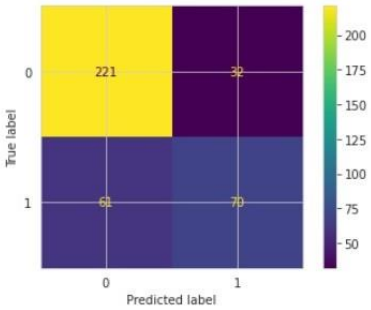
1) Naive Bayes Classifier



Gambar 4 Data Training 80% dan Testing 20%

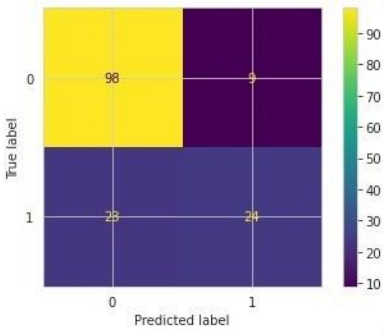


Gambar 5 Data Training 67% dan Testing 33%

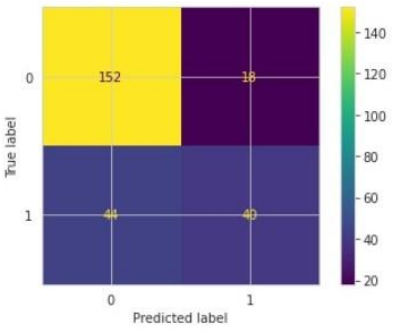


Gambar 6 Data Training 50% dan Testing 50%

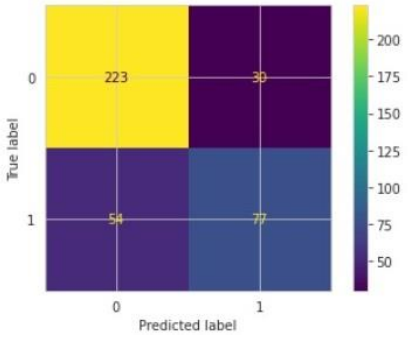
2) Random Forest



Gambar 7 Data Training 80% dan Testing 20%



Gambar 8 Data Training 63% dan Testing 37%



Gambar 9 Data Training 50% dan Testing 50%

G. Perbandingan Akurasi

Perbandingan akurasi dari klasifikasi *Naïve Bayes* dan *Random Forest* dapat dilihat Tabel 9.

TABLE IX. PERBANDINGAN AKURASI NAÏVE BAYES DAN RANDOM FOREST

No	Split data Training dan Testing	Akurasi	
		Naive Bayes Classifier	Random Forest
1	***Training 80% dan Testing 20%***	77%	79%
2	***Training 67% dan Testing 33%***	75%	76%
3	***Training 50% dan Testing 50%***	76%	78%
∑(rata-rata)		76%	77,67%

V. KESIMPULAN

Dalam penelitian ini, dilakukan perbandingan akurasi pendeteksian penyakit diabetes dengan teknik klasifikasi naïve bayes dan random forest dengan hasil penelitian yang menunjukkan bahwa berdasarkan 3 percobaan yang telah dilakukan kita dapat mengetahui nilai rata-rata dari akurasi naïve bayes dan random forest, dengan naïve bayes sebesar 76% dan random forest sebesar 77,67% sehingga dapat disimpulkan bahwa random forest memiliki akurasi lebih baik daripada naïve bayes dengan perbedaan 1,67%.

DAFTAR PUSTAKA

[1] WHO, “World diabetes day,” Atencion Primaria, 2015. http://www.who.int/diabetes/wdd_2015/en/.

[2] J. Apostolakis, “An introduction to data mining, Structure and Bonding,” 2010, doi: 10.1007/430_2009_1.

[3] P. M. Kellstedt and G. D. Whitten, Data Mining: Concepts and Techniques : Concepts and Techniques. 2018.

[4] F. Gorunescu, Data Mining: Concepts, models and techniques. 2011.

[5] H. Hairani, G. S. Nugraha, M. N. Abdillah, and M. Innuddin, “Komparasi Akurasi Metode Correlated Naive Bayes Classifier dan Naive Bayes Classifier untuk Diagnosis Penyakit Diabetes,” InfoTekJar (Jurnal Nas. Inform. dan Teknol. Jaringan), vol. 3, no. 1, pp. 6–11, 2018, doi: 10.30743/infotekjar.v3i1.558.

[6] S. Syarli and A. Muin, “Metode Naive Bayes Untuk Prediksi Kelulusan (Studi Kasus: Data Mahasiswa Baru Perguruan Tinggi),” J. Ilm. Ilmu Komput., vol. 2, no. 1, pp. 22–26, 2016.

[7] Y. Ramdhani, “Komparasi Algoritma LDA Dan Naïve Bayes Dengan Optimasi Fitur Untuk Klasifikasi Citra Tunggal Pap Smear,” Informatika, vol. II, no. 2, pp. 434–441, 2015, [Online]. Available: https://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/view/130%0Ahttps://ejournal.bsi.ac.id/ejurnal/index.php/ji/article/download/130/105

[8] M. Reza, S. Miri, and R. Javidan, “A Hybrid Data Mining Approach for Intrusion Detection on Imbalanced NSL-KDD Dataset,” Int. J. Adv. Comput. Sci. Appl., vol. 7, no. 6, pp. 1–33, 2016, doi: 10.14569/ijacsa.2016.070603.

[9] T. N. Nuklianggraita, A. Adiwijaya, and A. Aditsania, “On the Feature Selection of Microarray Data for Cancer Detection based on Random Forest Classifier,” J. Infotel, vol. 12, no. 3, pp. 89–96, 2020, doi: 10.20895/infotel.v12i3.485.

[10] Gde Agung Brahmana Suryanegara, Adiwijaya, and Mahendra Dwifebri Purbolaksono, “Peningkatan Hasil Klasifikasi pada Algoritma Random Forest untuk Deteksi Pasien Penderita Diabetes Menggunakan Metode Normalisasi,” J. RESTI (Rekayasa Sist. dan Teknol. Informasi), vol. 5, no. 1, pp. 114–122, 2021, doi: 10.29207/resti.v5i1.2880.

[11] “Diabetes.” [Online]. Available: https://www.kaggle.com/datasets/akshaydattatraykhare/diabetes-dataset?resource=download